Chemogenomic Analysis of the Druggable Kinome and Its Application to Repositioning and Lead Identification Studies

Ravikumar, Balaguru

2019-11-21

Ravikumar , B , Timonen , S , Alam , Z , Parri , E , Wennerberg , K & Aittokallio , T 2019 , ' Chemogenomic Analysis of the Druggable Kinome and Its Application to Repositioning and Lead Identification Studies ' , Cell chemical biology , vol. 26 , no. 11 , pp. 1608-1622 . https://doi.org/10.1016/j.chemb

http://hdl.handle.net/10138/321778
https://doi.org/10.1016/j.chembiol.2019.08.007

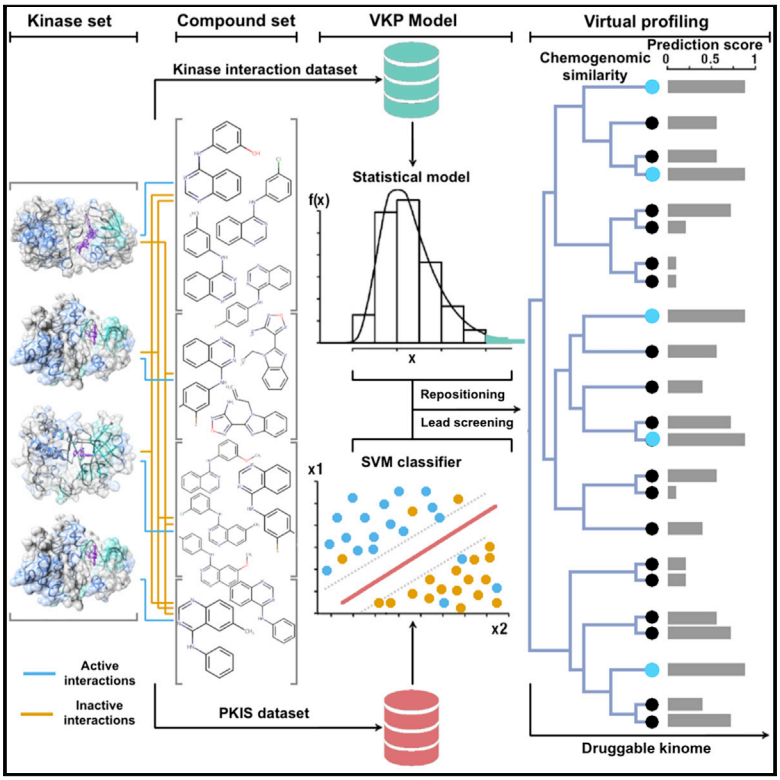
draft

Resource

Cell Chemical Biology

Chemogenomic Analysis of the Druggable Kinome and Its Application to Repositioning and Lead Identification Studies

Graphical Abstract



Highlights

- VKP prioritizes potent compound-kinase interactions for experimental validation
- VKP efficiently summarizes the chemogenomic landscape of druggable kinome
- VKP is widely applicable to both drug repositioning and lead identification studies
- The web tool enables profiling of compounds across 248 kinases simultaneously

Authors

Balaguru Ravikumar, Sanna Timonen, Zaid Alam, Elina Parri, Krister Wennerberg, Tero Aittokallio

Correspondence

tero.aittokallio@helsinki.fi

In Brief

The virtual kinome profiling (VKP) platform uses compound-kinase interaction information to prioritize potent activities for further pre-clinical evaluation. The platform uses the chemogenomic relationships of kinases to expedite the kinase inhibitor screening process, as demonstrated by several case examples. The platform and the accompanying datasets are implemented as a one-click web tool.

Ravikumar et al., 2019, Cell Chemical Biology *26*, 1–15
November 21, 2019
https://doi.org/10.1016/j.chembiol.2019.08.007

CellPress

Please cite this article in press as: Ravikumar et al., Chemogenomic Analysis of the Druggable Kinome and Its Application to Repositioning and Lead Identification Studies, Cell Chemical Biology (2019), https://doi.org/10.1016/j.chembiol.2019.08.007

Cell Chemical Biology
Resource

Chemogenomic Analysis of the Druggable Kinome and Its Application to Repositioning and Lead Identification Studies

Balaguru Ravikumar,[1] Sanna Timonen,[1] Zaid Alam,[1] Elina Parri,[1] Krister Wennerberg,[1,2] and Tero Aittokallio[1,3,4,*]
[1]Institute for Molecular Medicine Finland, FIMM, University of Helsinki, Helsinki 00014, Finland
[2]Biotech Research and Innovation Centre, University of Copenhagen, Copenhagen 2200, Denmark
[3]Department of Mathematics and Statistics, University of Turku, Turku 20014, Finland
[4]Lead Contact
*Correspondence: tero.aittokallio@helsinki.fi
https://doi.org/10.1016/j.chembiol.2019.08.007

SUMMARY

Owing to the intrinsic polypharmacological nature of most small-molecule kinase inhibitors, there is a need for computational models that enable systematic exploration of the chemogenomic landscape underlying druggable kinome toward more efficient kinome-profiling strategies. We implemented VirtualKinomeProfiler, an efficient computational platform that captures distinct representations of chemical similarity space of the druggable kinome for various drug discovery endeavors. By using the computational platform, we profiled approximately 37 million compound-kinase pairs and made predictions for 151,708 compounds in terms of their repositioning and lead molecule potential, against 248 kinases simultaneously. Experimental testing with biochemical assays validated 51 of the predicted interactions, identifying 19 small-molecule inhibitors of EGFR, HCK, FLT1, and MSK1 protein kinases. The prediction model led to a 1.5-fold increase in precision and 2.8-fold decrease in false-discovery rate, when compared with traditional single-dose biochemical screening, which demonstrates its potential to drastically expedite the kinome-specific drug discovery process.

INTRODUCTION

The human kinome encompasses a diverse array of molecular kinases with crucial regulatory and cellular functions (Manning et al., 2002), deregulation of which contributes to many complex diseases, especially cancers (Zhang et al., 2009). Protein kinases currently form the largest group of therapeutic targets for molecularly targeted anticancer drug treatment (Sun et al., 2017). A recent collective analysis of the target spectrum of various kinase inhibitors (KIs) detailed the proportion of clinically viable kinases in the kinomic landscape that is druggable, hereinafter referred to as the "druggable" kinome (Klaeger et al., 2017). A substantial fraction of KIs targeting the druggable kinome are ATP-competitive inhibitors (type I and type II KIs) that are known to display a higher degree of promiscuity compared with allosteric binders (type III and type IV KIs) (Munoz, 2017). This polypharmacological disposition among KIs is predominantly attributed to the residue conservation among the ATP-binding pockets of molecular kinases (Chen et al., 2007). Such promiscuity contributes to both therapeutic and adverse effects, making it imperative to maintain a stringent efficacy/safety ratio in anticancer drug development strategies (Ravikumar and Aittokallio, 2018). Comprehensive knowledge of the chemogenomic space underlying the druggable kinome is therefore critical for improving the success rates in clinical drug development phases (Fedorov et al., 2010).

Numerous kinome-wide target profiling studies have been carried out to explore both the cross-reactive potency of KIs and to elucidate their mechanism of action (Davis et al., 2011; Elkins et al., 2016; Fabian et al., 2005; Georgi et al., 2018; Metz et al., 2011). Such high-throughput kinome screening studies can also be outsourced, without expensive instrumentation requirements, through the commercially ventured kinase-profiling services, such as *SelectScreen* and *KINOMEscan* (Miduturu et al., 2011). Apart from profiling techniques, the emphasis on the identification of selective KIs has led to the development of dedicated kinome-specific chemogenomic screening libraries (Drewry et al., 2017; Elkins et al., 2016; Jones and Bunnage, 2017). Despite these developments, many of the new drug indications are identified through exhaustive screening experiments, rather than rational approaches, and such discoveries are largely dependent on the compound libraries subjected to biochemical profiling studies (Pemovska et al., 2015). Scaling up of compound libraries, although feasible, incurs additional time and cost constraints intrinsic to the phenotypic screening process. Furthermore, the inherent diversity of screening protocols and libraries has led to heterogeneous bioactivity profiles, which pose significant challenges for data integration procedures, hindering the data reuse in lead identification and drug repositioning studies (Arrowsmith et al., 2015; Orchard et al., 2011; Tang et al., 2018).

Computational models have been proposed and used as a cost-effective alternative to accelerate the drug discovery process (Lavecchia and Cerchia, 2016; Ravikumar and Aittokallio, 2018). A seminal work was the development of the compound-centric similarity ensemble approach (SEA). SEA underpins a

statistical framework that describes the chemoinformatic space of molecular targets by evaluating the similarity among active ligand sets of individual targets, thereby aiding in identification of novel drug indications and in deconvoluting the target landscape of compounds (Keiser et al., 2007, 2009). However, the wide generalizability of such model may impair their sensitivity, especially when focusing on kinase space, which sequentially has given rise to various kinome-specific prediction models (Christmann-Franck et al., 2016; Cichonska et al., 2017; Lo et al., 2018; Merget et al., 2017). More recently, various neural network and deep-learning models have been proposed for bioactivity prediction, some of which have outperformed the conventional prediction models (Koutsoukas et al., 2017; Ozturk et al., 2018). However, accurate predictions from such models often rely on computationally intensive algorithms, which may hinder their accessibility and wide usability by chemical biologists, unless implemented as stand-alone or web applications. Another fundamental limitation of most prediction models is their requirement of specific bioactivity data and structural information for model development. Although the integration of diverse datasets can enhance the performance of the prediction models (Iorio et al., 2010), it also significantly limits their applicability, especially when the aim is to systematically profile the full kinome space.

To address these limitations, we developed a model-guided kinome-profiling platform as an efficient means to analyze the chemogenomic landscape of the druggable kinome space. The aim was to systematically prioritize potent compound-kinase interactions for further biochemical and pre-clinical evaluation. The KI sets collected and used in this study form the most comprehensive kinome-specific compound resource used in a predictive analysis framework to date. The computational analysis platform integrates an enhanced statistical model with an efficient classifier developed using an ensemble support vector machine (eSVM) algorithm. We applied the platform to systematically explore the chemogenomic similarities among molecular kinases and to provide useful insights into potent compound-kinase associations. In addition to comparing the chemogenomic approach with the conventional sequence-based approach, we also performed a comparative analysis against similar statistical approaches (Keiser et al., 2007; Lin et al., 2013; Wang et al., 2016a). We computationally validated the performance of the platform by predicting the bioactivity classes in the published kinase chemogenomic sets (KCGS) (Drewry et al., 2017; Elkins et al., 2016). Finally, we demonstrate the wide applicability of our analysis platform in various kinome-specific lead identification and compound repositioning studies by experimentally validating the model predictions using biochemical assays. The compiled data resource and the accompanying computational model are deployed as a virtual kinome profiler web-application (https://virtualkinomeprofiler.fimm.fi/), to expedite the kinome-specific drug discovery process.

RESULTS

Chemogenomic Landscape of the Druggable Kinome

To develop and test the efficacy of our platform, we first curated a wide bioactivity spectrum of pharmacologically diverse KIs. The bioactivity profiles of these KIs includes information of their on- and off-target interactions, compound structural description, and target annotations, which were compiled and standardized from various public data resources (Bento et al., 2014; Tym et al., 2016). The resulting bioactivity dataset consisted of 248 kinases that cover 48% of the human kinome, and 74,033 KIs, which together form the current space of the druggable kinome in our study (see also Table S1). The affinity and potency measures of the ~250,000 compound-target interactions from the curated dataset were classified based on the activity threshold of 1 μM. The categorized active and inactive compound sets for each kinase in our panel cover all the key kinase sub-families (Figure 1A). To enumerate the chemogenomic association among various kinase targets, we calculated the active compound set similarities and inactive compound set dissimilarities using the Tanimoto and Dice coefficients (Fligner et al., 2002) (Equations 2, 3, 4, and 5; see the STAR Methods). To enhance the established SEA framework (Lin et al., 2013; Wang et al., 2016a), which uses merely ECFP4 fingerprints (FPs), we calculated the similarity and dissimilarity scores among ligand sets using eight different topological representations of compound FPs (Daylight, MACCS, ECFP4/6, Torsion, Atompairs, and FCFP4/6) (Figure 1B). Each molecular FP captures distinct facets of chemogenomic associations that exist among the analyzed kinome. The statistical significance (E values) of the similarity and dissimilarity scores among kinase targets were computed by comparing the observed score against a reference null distribution generated by randomly sampling a range of ligand set sizes (ranging from 100 to 1×10^6) from the curated dataset (Figure 2; Table S2). The E value statistic estimates the balanced likelihood of observing the score by random chance by accounting for the disproportionate compound set sizes related to each kinase targets (Equations 9, 10, 11, 12, and 13; see the STAR Methods). The statistical model therefore describes the chemogenomic landscape of kinases by accounting for the set size biases.

To systematically portray the chemogenomic links among the 248 kinase targets, we performed a comparative analysis of our chemical similarity-based approach against the traditional sequence similarity-based analysis of kinases (Figures 3A; see also S2). The sequence information and the associated kinase sub-family annotations were retrieved from the UniProt database (UniProt ConsortiumT, 2018) (Table S1). The sequence similarity of the kinases was estimated by querying a custom database of retrieved kinome sequences using the BLASTp alignment algorithm (Altschul et al., 1990). The E values estimated with the sequence-based BLASTp algorithm and from our chemogenomic statistical model using ECFP4 FPs were transformed to a distance matrix and subjected to unsupervised hierarchical clustering (see the STAR Methods). To distinguish the strength of target-target associations highlighted by these two orthogonal approaches, we performed a differential analysis of the obtained E value metrics (Figure 3B). This analysis identified certain kinome associations, such as FLT4 with RPS6KB1 and RIPK1, which were strongly related in their chemical space although being distant in their sequence similarities (Figure 3C, inset). Such associations highlight the added value of the chemogenomic approach when exploring the druggability of kinases that are dissimilar on a sequence level. Furthermore, we

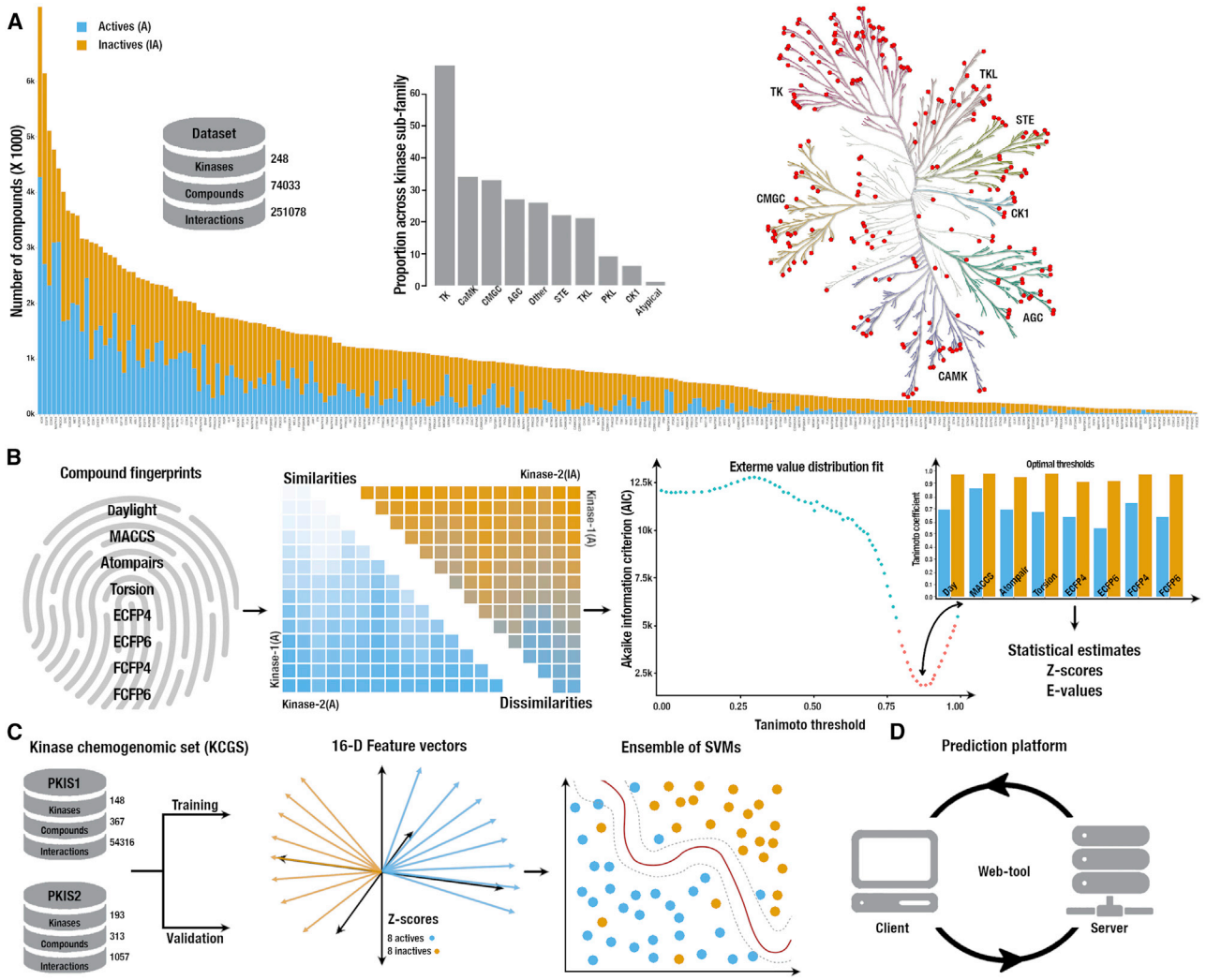


Figure 1. A Schematic Illustration of the Computational Framework Implemented for Chemogenomic Analysis and Virtual Profiling for the Druggable Kinome

(A) An overview of the kinase targets that form the druggable kinome panel in the study: the enumeration of active (A) and inactive (IA) compounds associated with each of the 248 kinase targets and their distribution across key kinase sub-families is represented in the kinome tree (see also Table S1).

(B) Statistical model implementation for each kinase target: eight compound fingerprints (FPs) were used in the study to calculate similarities and dissimilarities for the active and inactive compound sets (e.g., kinase-1 and kinase-2 in the heatmap), respectively, followed by fitting a generalized extreme value Gumbel distribution to find the optimal Tanimoto threshold (*Tc**) for each FP (see also Figures 2C and 2D). These thresholds were used to calculate the *Z* score and E value statistical estimates for compound activity class predictions. The *Z* scores for the similarity and dissimilarity calculations from the statistical model were used as input features in the SVM classification model.

(C) Machine-learning classification model implementation and testing in KCGS compound sets: PKIS datasets were used to train, test, and validate the ensemble of support vector machine (eSVM) classification model, which integrates the multiple FP feature vectors to ascertain the activity classes of compounds.

(D) The prediction platform was implemented as a web-application to facilitate kinome-profiling studies.

explored various similarity maps using an array of compound FP types, each revealing a distinct representation of the chemical landscape that exists in the kinome (see Figure S1). To elucidate the extent of diversity between sub-clusters of these similarity maps, we used the predetermined sub-family classification of protein kinases as a reference (Table S1), based on which we computed the adjusted Rand index and cophenetic correlation measures to quantify how accurately different FPs enable one to reconstruct the sub-family classes of kinases (Figure 3D). Although the statistical scores from FPs of active sets were correlated, as expected, each statistical model based on the unique FP features encodes complementary information of the chemogenomic spectrum underlying the druggable kinome.

Elucidating Binding Classes of KCGS Compound Sets

Based on the above results, we hypothesized that the statistical estimates (*Z* scores) obtained from the compound set similarity and dissimilarity scores based on the eight FPs could be used as a multifaceted feature panel for kinases. Such a panel captures essential information for predicting the binding classes

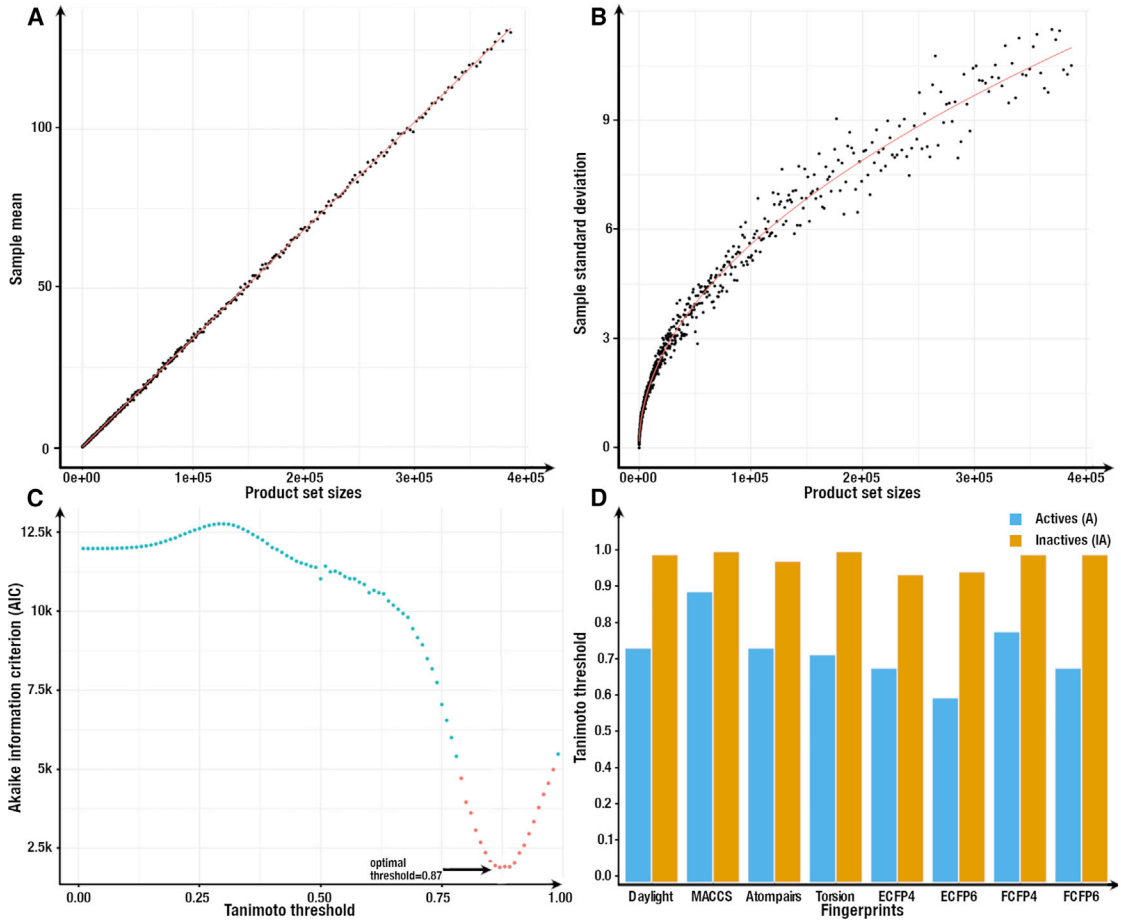


Figure 2. Statistical Model Parameter Estimation and Optimal Threshold Fits across Fingerprints

(A and B) The mean (A) and standard deviation (B) plot of original scores as a function of sampled set sizes for the reference null distribution when estimating active ligand set similarities using the MACSS FPs at the Tanimoto threshold of 0.87, their respective fits are indicated in red.

(C) The Akaike Information Criterion values obtained when fitting the *Z* score distributions for active ligand set similarities to a generalized Gumbel EVD across the 99 Tanimoto thresholds for the MACSS FPs, wherein the optimal threshold (*Tc**) was found to be 0.87 (see the STAR Methods).

(D) The optimal thresholds for various FPs used in the statistical model in enumerating both actives (similarities) and inactives (dissimilarities) feature vectors (see also Table S2).

(active versus inactive) of a compound for a given kinase target. To effectively integrate the non-redundant information from such 16-dimensional (8 actives and 8 inactives) feature vectors, we implemented a classification algorithm based on our statistical chemogenomic model (Figures 1C and 3D). To exclude any inadvertent impacts from heterogeneous screening methodologies when developing and validating our classification algorithm, we used the bioactivity profiles from published kinase chemogenomic studies (Drewry et al., 2017; Elkins et al., 2016) that use the Nanosyn technology and DiscoverX KINOMEscan protocols, respectively. The published kinase inhibitor sets (PKIS1 and PKIS2) and their interaction measures were used as the training and independent held-out validation datasets, that is, they were blinded when developing the statistical model (Figure 1C). The compound-kinase interaction profiles from the KCGS screening studies were assigned to binary reference classes (positives or negatives) using a threshold of 1 μM (see also Figures S3A and S3B). Before implementing the classification algorithm, we evaluated the predictive power of the individual statistical attributes (E value estimates from the various FPs) in distinguishing the binding classes in the PKIS1 dataset, where the predicted classes were defined by applying an E value threshold of 1×10^{-10} (Figure S4A). Based on the Fβ performance metric across selected baseline machine-learning (ML) models (Equations 14, 15, 16, 17, 18, and 19; see the STAR Methods), we identified a non-linear SVM classifier using a radial basis function (RBF) kernel having the highest classification performance in our study (Figure S4B).

We implemented a 5-fold cross-validation testing protocol within the PKIS1 dataset to train and fine-tune the optimal hyperparameters (C and Gamma) of the SVM-RBF model (see Figure S4C and S4D). An impending characteristic in classifying compound-kinase bioactivity profiles is the inherent data imbalance between the class labels (the average ratio of positives to negatives in PKIS1 is 1:20; Figure S3A). To account for such imbalanced classification task, we adopted an ensemble schema, termed ensemble of under-sampled SVM (eSVM), which has been shown to improve the performance and

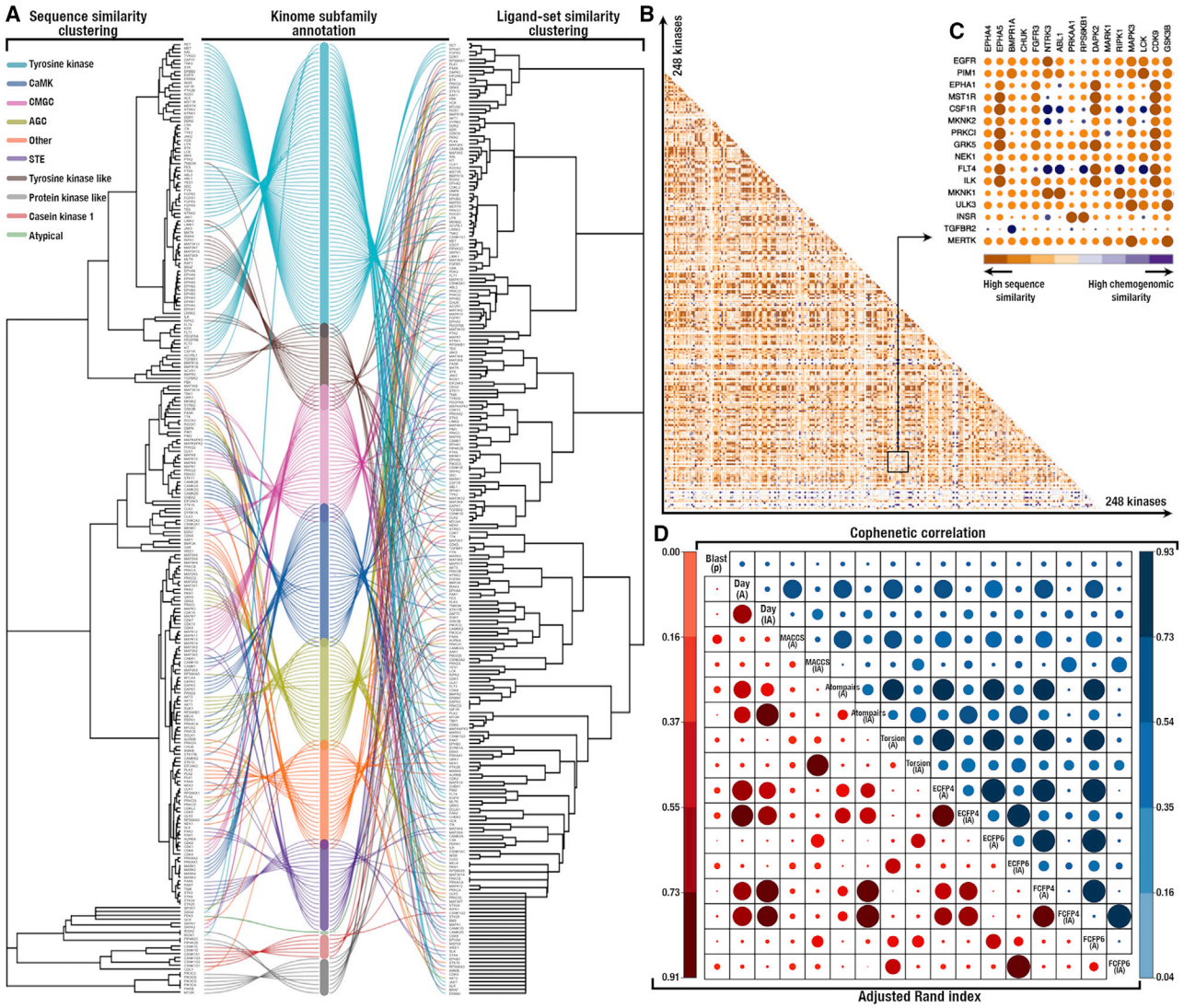


Figure 3. Analysis of Chemogenomic Landscape Underlying the Druggable Kinome

(A) Comparative analysis of kinase-kinase similarities based on the sequence similarity space and the chemical similarity space among the 248 kinase targets, where the former was calculated using the BLASTp algorithm and the latter using ECFP4 FP (the kinase sub-families are color coded; a high-resolution version of the cluster solutions is given in Figure S2). The cosine distances between E values estimated based on these orthogonal approaches were used to depict kinase-kinase association in the dendrograms.

(B and C) A differential heatmap, highlighting the strength of those associations reflected by the two approaches (B). A subset of the heatmap comparing the targets associated with stronger sequence similarity in comparison with chemogenomic similarity (C) (inset) (see the STAR Methods).

(D) The adjusted Rand index and cophenetic correlation metrics calculated based on similarity and dissimilarity statistical estimates among the active (A) and inactive (IA) compound sets, respectively, across the different FPs. The FPs capture non-redundant information of the chemogenomic representations among kinases (the chemogenomic clustering solutions based on the different compound FPs excluding ECFP4 are shown separately (see also Figure S1).

generalizability of the SVM classifier (Kang and Cho, 2006) (see the STAR Methods). All of the 54,316 investigated compound-kinase interactions among 367 compounds and 148 kinases were assigned to an activity class by aggregating the prediction outcomes from 23 individual classifiers through a majority voting system; similarly, all the performance metrics of the eSVM model were evaluated by averaging the measures over the 23 independent ensembles (Equations 14, 15, 16, 17, 18, and 19; see the STAR Methods). The eSVM ensemble model classified the PKIS1 dataset with a precision rate of 84.8% and an AUC of 0.74, and the Fβ metric accounting both for precision and recall was 0.72. Comparison of the AUC metric of the trained eSVM model with that of individual FP measures highlights the improvements gained by the integrated eSVM model when predicting the true activity classes of the KIs ($p < 10^{-10}$, DeLong's test; Figure 4A). Furthermore, when recapitulating the activity classes in the external PKIS2 validation dataset (a total of 1,057 interactions; Figure S3B), the measured equilibrium dissociation constant (K_D) values of the interactions classified as actives by the eSVM model showed a median K_D level of 0.093 μM, a

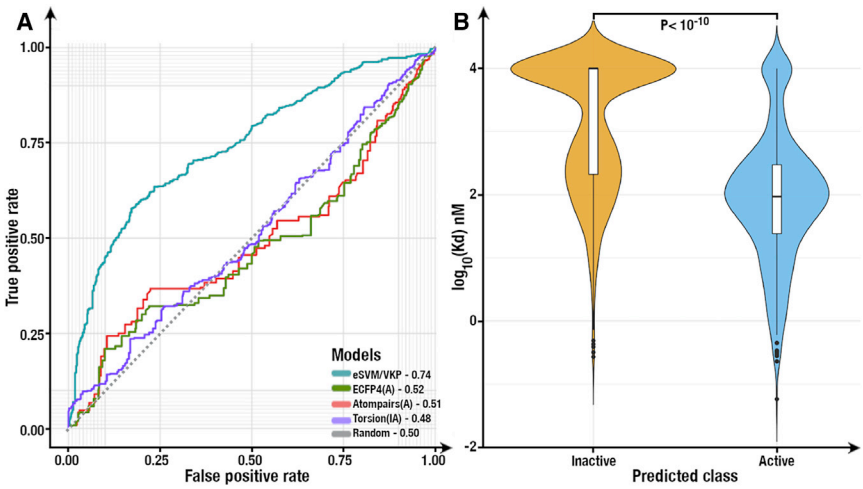


Figure 4. Evaluation of the Ensemble SVM-RBF Model

(A) The receiver operator characteristic (ROC) curve comparing the performance of the integrated ensemble SVM classifier (eSVM) with the statistical E value estimates from the individual FPs (ECFP4-A; Atompairs-A; Torsion-IA; active [A] and inactive [IA] compound sets) in the PKIS1 dataset (Elkins et al., 2016). The area under the ROC curve (AUROC) for the various models are provided in the color-coded legend (the differences of the observed AUROCs were highly significant; $p < 10^{-10}$, DeLong's test).

(B) The distribution of the bioactivity values of the active and inactive classes in the independent PKIS2 validation dataset (Drewry et al., 2017) as predicted by the eSVM model. The difference between the two classes was highly significant ($p < 10^{-10}$, Wilcoxon test).

significantly higher affinity than those classified as inactives ($p < 10^{-10}$, unpaired two-sided Wilcoxon test at 99% CI; Figure 4B). These results suggest an accurate predictive power of our systematic computational approach in elucidating compound-target interactions across diverse kinase sets of the druggable kinome (Figure S3C). We also observed a 9.1-fold enrichment of true hits when comparing the model predictions against the experimentally defined active class ($K_D < 1$ μM) in the PKIS2 dataset ($p < 10^{-10}$, Fisher's exact test).

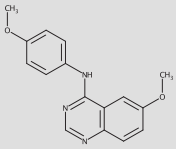
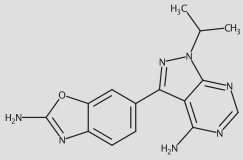
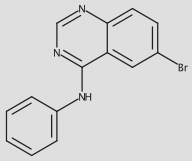
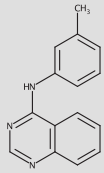
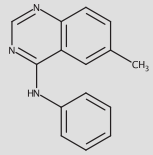
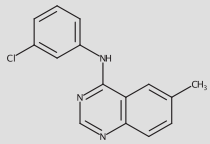
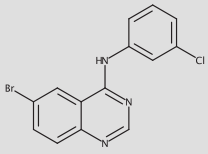
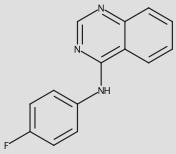
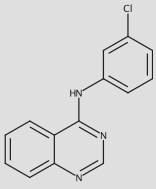
Applications to Repositioning and Lead Screening

After validating the competence of our computational platform in the PKIS datasets, we next extended its application to predicting potent small-molecule KIs. We addressed two prevalent tasks in the pre-clinical drug discovery process, namely, compound repurposing and lead molecule identification. The repurposing collection consisted of compounds and their target annotations, compiled and curated from multiple data resources such as the Drug Repositioning Hub, DrugBank, and IUPHAR (Corsello et al., 2017; Harding et al., 2018; Wishart et al., 2018); this collection of 18,077 compounds comprises mostly non-KIs and KIs whose activity profile is poorly established (see Table S3). Similarly, the lead molecule library is a compendium of 133,631 potential lead molecules obtained from various chemical diversity libraries (including Specs consortium, Tripos collection, MicroSource Spectrum, ChemBridge DIVERSet, and ChemDiv diversity collections; Table S3). Before activity class prediction, the compounds from both of these libraries were curated, among other criteria, by selecting small-molecule inhibitors with molecular weights <700 Da (see the STAR Methods). The activity classes of these 151,708 compounds were predicted across the 248 kinases in our druggable kinome panel, resulting in approximately 37 million kinase-compound associations, using the eSVM-RBF model.

In the current target profiling study, given an instance of compound-kinase interaction, the practical prediction task involves estimating a statistical measure of similarity and dissimilarity of the query compound with the predefined active and inactive ligand sets of each kinase, to make the decision whether or not the particular compound should be considered as potent inhibitor of the selected kinase. To improve the sensitivity of the activity class predictions, we further aggregated the statistical estimates (*Z* scores) from the FPs and classified the compound-kinase activity using the eSVM-RBF classifiers. Applying an objective threshold of ≥ 0.875 for both normalized eSVM score and for the prediction score (Equations 7 and 8; STAR Methods), defined based on a pilot study among well-established kinase targets (see Table S4), we selected 51 compound-kinase interactions between 51 compounds and 8 kinases for further biochemical validation (Table S5). We did not favor any specific targets in the experimental validations, rather selected the targets based solely on the predicted potency as estimated by the data-driven computational models. Among the 51 selected interactions, 46 were novel associations and 5 compound-kinase interactions have previously been reported, thereby serving as positive controls of our model predictions (Table S6).

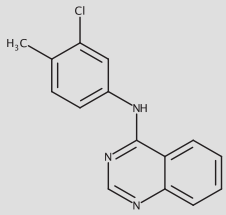
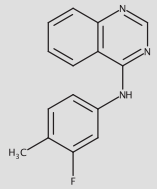
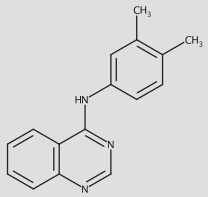
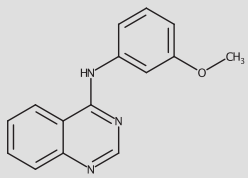
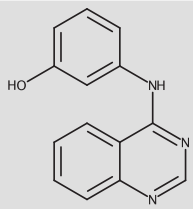
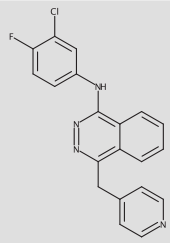
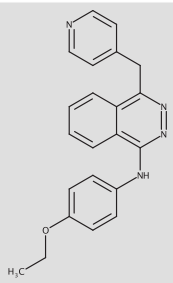
The experimental validations of the model predictions were performed sequentially using a two-phase screening procedure, namely, a single concentration assay followed by a dose-response assay, both of which use a cell-free biochemical kinase testing protocol (Equation 1; see the STAR Methods). Those compound-kinase interactions (37% of the 51 predicted interactions) that exhibited a residual activity of 50% or less in the single concentration assay (10 μM of compound) were selected for the dose-response study (12 concentration doses ranging from 0.0375 nM to 12.5 μM) to estimate the half-maximal inhibitory concentration (IC_{50}) values (Tables 1; see also S5). The potency values of the five known compound-kinase associations (compounds 2, 6, 7, 11, and 23) were similar to those established in previous studies (Bamford et al., 2005; Kolb et al., 2008; Metz et al., 2011; Wang et al., 2016b), supporting the accuracy of the model and the consistency of our experimental assay (Table S6). When examining the novel compound predictions validated from the repurposing collection, we identified LY456236, a selective, non-competitive metabotropic glutamate receptor 1 (mGlu1) inhibitor that has been shown to inhibit phosphoinositide hydrolysis *in vitro* ($IC_{50} = 0.145$ μM) (Shannon et al., 2005). Interestingly, our study confirmed that LY456236 inhibits also the epidermal growth factor receptor (EGFR) with an IC_{50} value of

Table 1. Compounds Predicted and Validated across Various Kinase Targets, with Their Single-Dose and Dose-Response Activities

Compound	Structure	Intended Target	Predicted Target (UniProt ID)	% Residual activity (10μM)	IC_{50} (μM)
LY456236		mGlu1	EGFR (UniProt: P00533)	16.6	0.918
Sapanisertib		mTOR	HCK (UniProt: P08631)	16.8	1.767
Compound 1		–[a]	EGFR (UniProt: P00533)	33.8	0.09
Compound 2		–	EGFR (UniProt: P00533)	12.3	0.104
Compound 3		–	EGFR (UniProt: P00533)	41.7	0.498
Compound 4		–	EGFR (UniProt: P00533)	8.7	0.03
Compound 5		–	EGFR (UniProt: P00533)	4.6	0.02
Compound 6		–	EGFR (UniProt: P00533)	38.8	1.573
Compound 7		–	EGFR (UniProt: P00533)	5,6	0.067

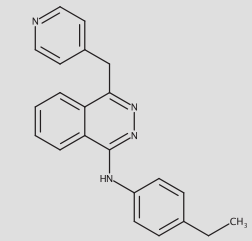
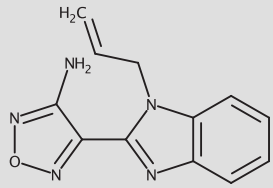
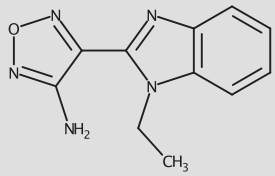
(*Continued on next page*)

Table 1. ***Continued***

Compound	Structure	Intended Target	Predicted Target (UniProt ID)	% Residual activity (10μM)	IC_{50} (μM)
Compound 8		–	EGFR (UniProt: P00533)	6.8	0.029
Compound 9		–	EGFR (UniProt: P00533)	21.8	0.285
Compound 10		–	EGFR (UniProt: P00533)	25.9	0.273
Compound 11		–	EGFR (UniProt: P00533)	18.6	0.388
Compound 12		–	EGFR (UniProt: P00533)	43.8	1.684
Compound 16		–	FLT1 (UniProt: P17948)	2.4	0.085
Compound 17		–	FLT1 (UniProt: P17948)	−0.3	0.242

(*Continued on next page*)

Table 1. *Continued*

Compound	Structure	Intended Target	Predicted Target (UniProt ID)	% Residual activity (10μM)	IC_{50} (μM)
Compound 18		–	FLT1 (UniProt: P17948)	−2.7	0.016
Compound 22		–	RPS6KA5/MSK1 (UniProt: O75582)	35.3	0.587
Compound 23		–	RPS6KA5/MSK1 (UniProt: O75582)	−5.5	0.118

[a]Compounds from lead molecule libraries are marked with the intended target as "–" (for the full list, see Table S5). The first two compounds represent repurposing cases. mTOR, mammalian target of rapamycin.

0.91 μM (Table 1; Figure 5A). Similarly, sapanisertib (TAK-228), a potent and selective mammalian target of rapamycin inhibitor (IC_{50} < 0.10 μM), was confirmed to inhibit also the HCK protein kinase with an IC_{50} value of 1.76 μM (Table 1, Figure 5A).

In addition to these possible repositioning opportunities, we also experimentally validated novel predictions spanning 17 potential lead molecule interactions across three different kinases from various kinome sub-families (compounds 1–23 in Tables 1 and S5). Of the 12 compounds tested for EGFR inhibition, we identified two compounds (compound 4 and compound 5) whose IC_{50} values were ≈0.02 μM, i.e., only 1 log-fold less potent than that of the control compound afatinib (IC_{50} = 0.005 μM; Table 1; see Figure S5A). When validating the predicted FLT1 inhibitors, we identified compounds 16 and 18, which exhibited IC_{50} values of 0.08 and 0.01 μM, respectively. Interestingly, these inhibitors were found to be more potent than the FLT1 control compound sunitinib (IC_{50} = 0.23 μM; Table 1; Figures 5B, see S5B). Similarly, when experimentally assessing the RPS6KA5 (MSK1) inhibitors, compounds 22 and 23 were found to inhibit MSK1 kinase with IC_{50} values of 0.58 and 0.11 μM, respectively, displaying a similar potency to that of sunitinib (IC_{50} = 0.55 μM, Table 1; Figure 5C).

The biochemical validation of these novel compound-kinase interactions demonstrates the wide applicability of our proposed approach as an efficient computational screening platform for the systematic profiling of the druggable kinome. The overall precision of the computational-experimental strategy was 84% (16/19 interactions with IC_{50} < 1 μM; Table 1), which provides a 1.5-fold increase in positive predictive value (PPV) and a 2.8-fold decrease in false-discovery rate (FDR), compared with the two-phase experimental assay (single and dose-response assay) used in the PKIS2 chemogenomic study (Drewry et al., 2017).

DISCUSSION

Most previous approaches to deconvoluting the kinase spectrum of anticancer drug treatments have been biased toward clinically validated kinase targets (Fedorov et al., 2010). Although developments in high-throughput screening techniques and recent initiatives in elucidating the druggable genome have enabled more holistic kinome-wide profiling strategies (Georgi et al., 2018; Miduturu et al., 2011; Rodgers et al., 2018), there is a need for systematic computational frameworks that enable cost-effective analysis and profiling of the chemogenomic landscape of the druggable kinome. This work introduces a systematic methodology to comprehensively investigate the chemogenomic space of the druggable kinome and provides an efficient prediction platform to identify potent kinome-specific activities (Figure 1A). Compared with an existing general statistical model, such as SEA, our specific focus was on kinome space. Furthermore, while SEA estimates similarities among active ligand sets of targets using ECFP4 compound FPs only (Lin et al., 2013), we extended the statistical approach by several key improvements, wherein both active ligands set similarities and inactive ligand set dissimilarities among kinase targets were estimated using 8 different molecular FPs, giving rise to 16 independent representations of the kinase chemogenomic space (Figures 1B and 2). We further combined these distinct features with an ensemble SVM classification model to capture the full chemical similarity space of the druggable kinome (Figure 1C). The implemented ligand-based approach identifies kinome associations that can

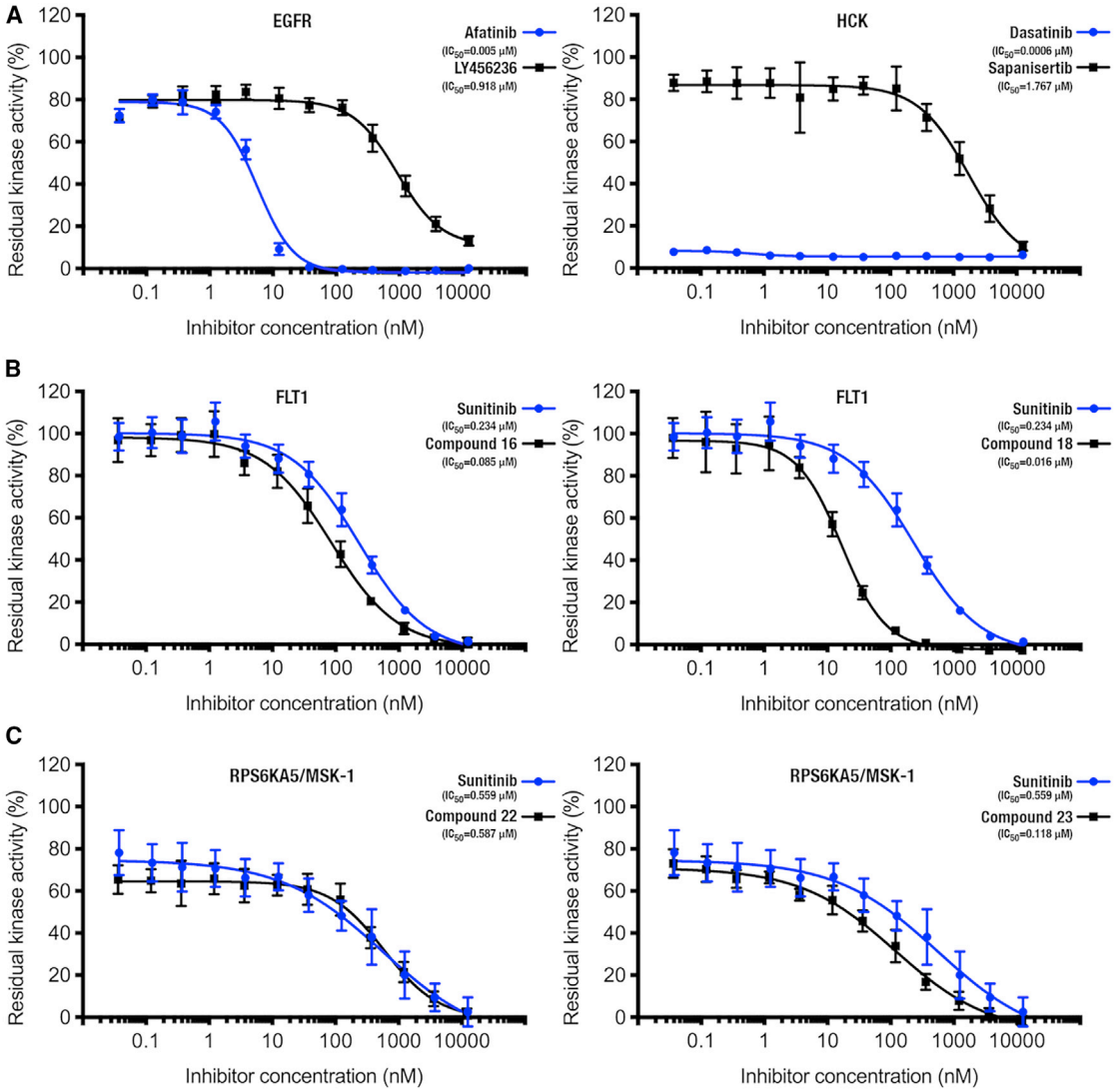


Figure 5. Dose-Response Assay for Compounds Predicted Among Repurposing and Lead Screening Libraries

(A) The dose-response curve of the predicted repurposing compounds LY456236 and sapanisertib against EGFR and HCK protein kinases, in comparison with their positive controls, afatinib and dasatinib, respectively.

(B) The dose-response curve of potential hit molecules (compounds 16 and 18) predicted to inhibit FLT1 kinase (the dose-response assay for compound 17 is shown [see also Figure S5B] in comparison with their positive control sunitinib).

(C) The dose-response curve of potential hit molecules (compounds 22 and 23) predicted to inhibit MSK-1 kinase, in comparison with their positive control, sunitinib. The IC_{50} dose-response measures for each compound-target interaction are provided in parentheses.

be distant in their sequence space but still strongly related in their chemical space (Figures 3A, see also S2). For instance, the chemogenomic similarity between FLT4 tyrosine kinase protein and RPS6KB1, which belongs to the AGC kinase sub-family, highlights the common scaffold similarities shared between these distant protein kinases, thereby aiding in multi-targeted drug designing (Besnard et al., 2012) (Figures 3B and 3C). Similar to previous studies, we observed a high degree of correlation (>0.42) among the similarity estimates from the active FPs (Hert et al., 2008; Wang et al., 2016a), whereas the additional dissimilarity estimates across inactive FPs showed a lower correlation (<0.57) (Figure 2D), which indicates the additional information obtained through this integrated approach. Furthermore, when categorizing the activity classes in the KCGS dataset (Figure S3A), the previously established protocol of merely using statistical estimates from individual FPs (e.g., E value thresholds in SEA) resulted in a lower performance compared with the eSVM model that aggregates information from multiple FPs from both active and inactive ligand sets (Figures 4A, see also S4A and S4B). The existing implementations of such statistical models have primarily focused on ranking the various compound representations in terms of FPs, and then selecting an optimal FP to perform the similarity estimation (Hert et al., 2008; Keiser et al., 2007; Wang et al., 2016a). By integrating statistical estimators for both active and inactive sets using an efficient ML algorithm, we demonstrate how the implemented computational platform

significantly reduces misclassification errors (Figure 4A). Similarly, in the external validation PKIS2 compound sets (Figure S3B), the active compounds predicted by the model showed a median bioactivity value of <0.10 μM (Figure 4B).

As an application of our computational framework, we profiled the kinome-centric interactions for a compendium of 151,708 compounds. Before investigating these distinct repurposing and drug-like molecule libraries, the compounds were curated using the pan assay interference compounds filter to exclude any unforeseen and non-specific interactions (Baell and Holloway, 2010). In aggregation, approximately 37 million compound-kinase associations were subjected to eSVM classification, and of the classified active binders, 51 interactions were later validated through an experimental assay (Table S5). Through subsequent biochemical assays, we identified 19 small-molecule inhibitors, several of which exhibited a potency of <1 μM in the dose-response assays, which is equal to the threshold used to designate active interactions in the statistical model (Table 1; see the STAR Methods). Apart from the two potential repositioning predictions (Figure 5A), we identified potent hit molecules for EGFR, MSK1, HCK, and FLT1 kinases, with activities similar to, or more potent than their respective control compounds used in the dose-response assays (Figures 5B and 5C; see also S5). Compared with standard high-throughput kinase-profiling experiments, with an approximate success rate of 18%, estimated based on the published data (Davis et al., 2011), when applying the same activity threshold of K_D < 1 μM), we can estimate the likelihood of obtaining 19 or more positive predictions in a random sample of 51 compound-kinase pairs (p = 0.00024, assuming a binomial distribution for the number of successes). The overall PPV of 84% (16/19) and FDR of 15% (3/19) for our computational-experimental approach provides 1.5-fold increase in PPV and 2.8-fold decrease in FDR, compared with the two-phase experimental assay implemented in PKIS2 (PPV = 57% and FDR = 42%), indicating that one could replace the single-dose initial screen with the computational platform to enrich the number of true actives among the predicted interactions, and, more importantly, to gain significant reductions both in time and costs of the experimental screening process.

In total, 63% of the predicted interactions (32 of the 51 tested) displayed a residual activity of >50% in the single-dose assay, and hence were categorized as false-positives and excluded from the further dose-response study (Table S5). In general, computational models performing such classification tasks can never be devoid of misclassification errors. We note that the optimal hyperparameters (C and Gamma) of the classification model were tuned using the Fβ metric with a β value of 0.5 (see the STAR Methods), hence giving more weight to precision than recall. Notably, most of the observed false-positives were in the predicted associations with kinase targets CDK6, TGFBR2, CHEK2, and FLT3 (Table S6). The *Z* score estimates from torsion FPs obtained for the true-positive interactions (EGFR, FLT1, MSK1, and HCK) and false-positive predictions (CDK6, TGFRB2, CHEK2, and FLT3) showed a significant difference (p = 0.03 for active and p = 0.0002 for inactive *Z* scores, respectively, unpaired two-sided Wilcoxon test at 99% CI; Figure 6A). We also observed a significant difference in compound diversity among the ligand sets for the false-positive target predictions, when compared with the diversity of compounds involved in the true-positive predictions (two-sample Kolmogorov-Smirnov test D = 0.195; $p < 10^{-10}$ at 99% CI; Figure 6B). The performance of any compound-centric prediction platform is bound to be significantly dependent on the size and extent of diversity observed among the active and inactive ligand sets accompanying the kinase targets. Therefore, the more diverse these compound sets are, the more distinct scaffolds will be identified and used by the model for virtual screening. Although the results in the PKIS datasets demonstrated the wide applicability of the computational platform to various kinases (Figure S3C), there was a clear enrichment for EGFR and TGFBR2 targets among the selected hits (Table S5), which is likely due to the chemical composition of the screened compound collection used in the present work. We therefore further analyzed the diverse scaffolds underlying the computational hits selected using Scaffold Hunter (Wetzel et al., 2009), which depicted 8 distinct maximum common substructures among the 51 selected compounds (Figures 6C and 6D), indicating various scaffolds for the compounds selected for experimental validation. In future research, the molecular scaffolds highlighting the commonalities of active binders across distinct kinase targets could also be used for multi-target *de novo* drug-design studies (Besnard et al., 2012). In addition to the standard compound FPs, the performance of the computational platform might be further improved by incorporating other pharmacophore-based features for activity class prediction, similar to the features highlighted in ColBioSFlavRC (Bora et al., 2016).

This study has certain limitations, the most prominent of which is the preference for implementing a compound-centric model. Although target-centric models, such as KinomeFEATURE, have recently been used for compound activity predictions (Lo et al., 2018), they are often constrained by the availability and quality of structural data essential for the comprehensive kinome-profiling studies. Similarly, even though augmenting other omics datasets to the prediction model might considerably enhance the classification performance (Iorio et al., 2010), as was shown, for instance, in the DEMAND algorithm (Woo et al., 2015), the requirement of such information further restricts the wider applicability and translatability of the prediction model. Currently, our compound-centric model only requires the structural description of the compounds (SMILES), whereas integrating other data resources would escalate the model's data-dependency requirements in performing similar classification tasks, therefore hindering its application to kinases or compounds lacking the required additional information. Existing *in silico* methodologies often implement various ML algorithms to facilitate the prediction of potent drug indications and target deconvolutions studies (Lavecchia, 2015). Newer methods include recently developed deep neural network models, such as DeepDTA, which uses one-dimensional representations of drugs and targets for bioactivity affinity predictions (Ozturk et al., 2018). Our study relied on an SVM algorithm for the classification task, and this selection was due to the "tried-and-true" effectiveness of SVMs when using the FP representations of the compounds as input features for similarity searches (Geppert et al., 2008). An additional benefit of the SVM model is its support for an ensemble learning approach, similar to that adopted in a previous study (Kang and Cho, 2006). Although classification and

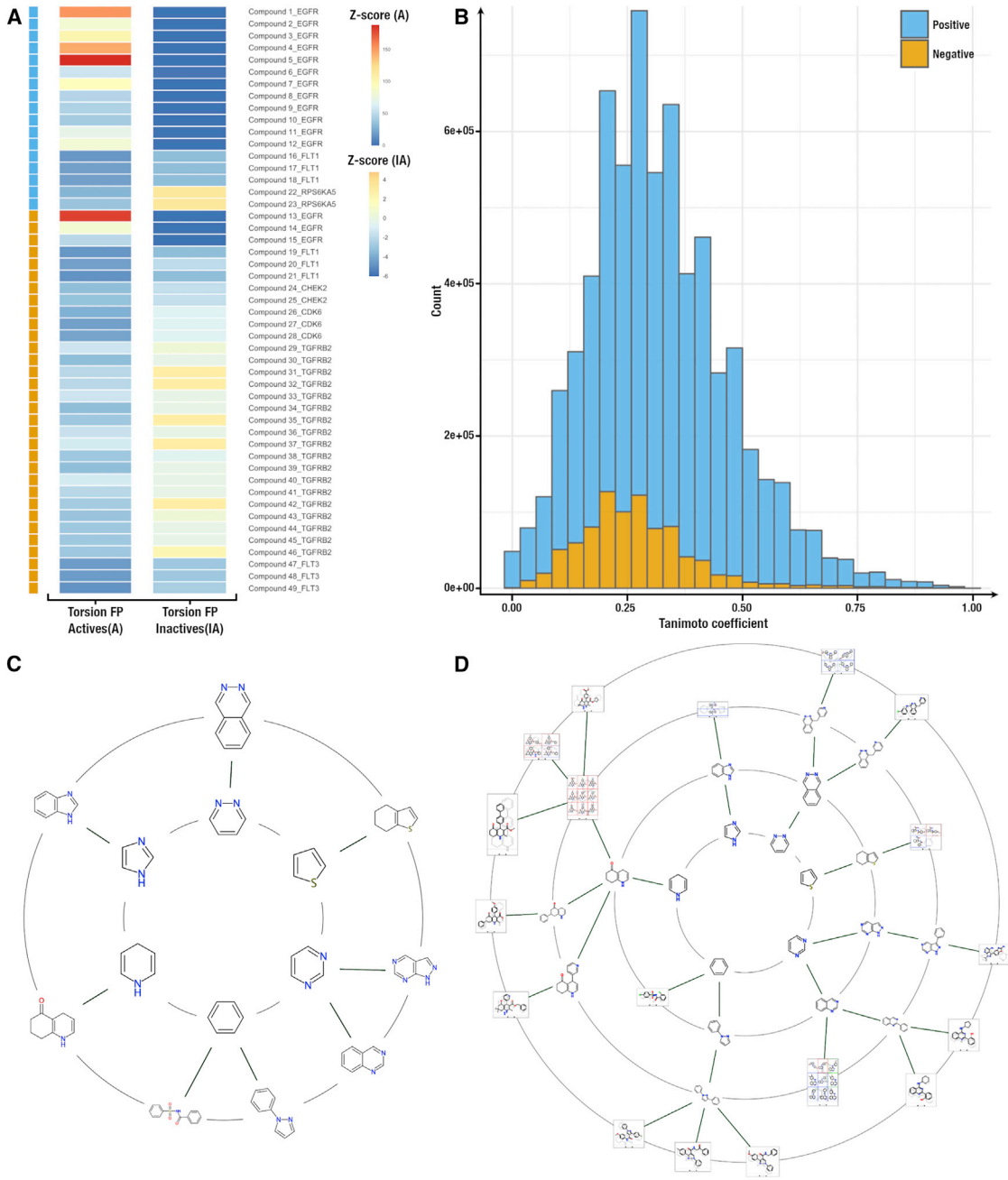


Figure 6. Comparing the Activity Classes and Scaffold Analysis of Assayed Interactions

(A) The *Z* score estimates for the 49 validated compound-kinase interactions obtained using torsion active (A) similarity and inactive (IA) dissimilarity estimates, where the true-positive (TP) predictions confirmed with biochemical assay are highlighted in Picton blue and the false-positives (FP) in orange peel. The difference in *Z* scores between the TP and FP predictions was significant for both A and IA sets ($p < 0.05$, unpaired two-sided Wilcoxon test).

(B) The chemical diversity of compound set similarity estimates obtained from the statistical model using torsion FPs across the targets validated as TPs (EGFR, FLT1, and MSK-1, Picton blue) and as FPs (TGFRB2, FLT3, CDK6, and CHEK2, orange peel). The difference in the distributions of the compound diversity was evaluated using the two-sample Kolmogorov-Smirnov test ($D = 0.195$; $p < 10^{-10}$).

(C) Scaffold analysis carried out using Scaffold Hunter v.2.6.0, illustrating the eight different maximum common substructures that underlie the molecular scaffolds of the 51 selected computational hits from virtual kinome profiling.

(D) The scaffold clustering and tree diagram that represents the subset of scaffolds and their sequential generation from MCS, corresponding to the 51 compounds selected for the experimental validation.

regression models may prioritize distinct structural features for activity prediction (Rodriguez-Perez et al., 2017), the choice of using a classification algorithm instead of a regression framework (Cichonska et al., 2017) stems from the eventual application of the platform as a pre-screening tool for prioritizing potent compound-kinase interactions, thereby reducing the number of compounds that are subjected to further biochemical or pre-clinical testing. Finally, the entire study was designed specifically for the druggable kinome, and such a restriction was enforced to expedite the kinome-specific drug discovery process. The model uses the high degree of polypharmacology observed among KIs (Hanson et al., 2018) to improve the specificity of compound-kinase interaction predictions compared with more generalized models. The application of the proposed platform is currently limited to the 248 kinases that form the druggable kinome panel. However, by adding other screening studies for kinase and other target classes (Fedorov et al., 2010; Rodgers et al., 2018), the current platform can be extended in coverage, and also to enable predictions for mutant kinase targets as well as for selectivity analysis. However, such more detailed prediction tasks require more comprehensive and standardized dose-response bioactivity data to allow for systematic and large-scale compound selectivity analyses, which are critical for KI discovery.

SIGNIFICANCE

The proposed chemoinformatic platform enables systematic exploration of the polypharmacological space of protein kinases. The various proof-of-concept case studies demonstrated the capability of the computational model in diverse drug discovery endeavors, ranging from investigating the compound's repositioning potential to implementing efficient lead identification strategies. These applications will significantly enhance the hit-lead optimization phase of a rational-based drug discovery process. Systematic mapping of the chemogenomic associations of kinases will further pave the way for more efficient multi-targeted drug development processes that maintain an acceptable efficacy/safety ratio for promiscuous KIs. Incorporation of our computational platform into traditional kinome-profiling campaigns should greatly accelerate the screening process and improve its accuracy. In addition, the implementation of the model as an easy-to-use web-application, which requires minimal compound description for the prediction task, provides chemical biologists with informed suggestions regarding their compounds' activity across the druggable kinome.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.chembiol.2019.08.007.

ACKNOWLEDGMENTS

We thank Drs. Suleiman Khan, Jing Tang, and Gopal Peddinti for their insightful suggestions for the work. We thank Olle Hansson at FIMM and Kimmo Mattila at CSC–IT Center for Science Ltd.—for setting up the computational environment necessary for this study. We thank the HTB unit of FIMM for providing their chemical collections and the various compound diversity panels used in this study. This work was supported by the Cancer Society of Finland grant to T.A. and K.W., the Sigrid Jusélius Foundation, Finland to T.A., grants from the Academy of Finland, Finland (292611, 279163, 295504, 310507, 313267, and 326238 to T.A.), and the doctoral fellowship from the Integrative Life Science (ILS) doctoral program to B.R.

AUTHORS CONTRIBUTIONS

B.R. and T.A designed the study. B.R. collected the data and developed the computational model. S.T. and E.P. performed the biochemical validations. B.R. and Z.A. implemented the web-application. T.A. and K.W. supervised the study and helped in the experimental design. B.R., S.T., K.W., and T.A. wrote the manuscript, which was revised and approved by all the authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 13, 2019
Revised: July 18, 2019
Accepted: August 21, 2019
Published: September 11, 2019

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Arrowsmith, C.H., Audia, J.E., Austin, C., Baell, J., Bennett, J., Blagg, J., Bountra, C., Brennan, P.E., Brown, P.J., Bunnage, M.E., et al. (2015). The promise and peril of chemical probes. Nat. Chem. Biol. *11*, 536–541.

Baell, J.B., and Holloway, G.A. (2010). New substructure filters for removal of Pan Assay Interference Compounds (PAINS) from screening libraries and for their exclusion in bioassays. J. Med. Chem. *53*, 2719–2740.

Bamford, M.J., Alberti, M.J., Bailey, N., Davies, S., Dean, D.K., Gaiba, A., Garland, S., Harling, J.D., Jung, D.K., Panchal, T.A., et al. (2005). (1H-Imidazo[4,5-c]pyridin-2-yl)-1,2,5-oxadiazol-3-ylamine derivatives: a novel class of potent MSK-1-inhibitors. Bioorg. Med. Chem. Lett. *15*, 3402–3406.

Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Kruger, F.A., Light, Y., Mak, L., McGlinchey, S., et al. (2014). The ChEMBL bioactivity database: an update. Nucleic Acids Res. *42*, D1083–D1090.

Besnard, J., Ruda, G.F., Setola, V., Abecassis, K., Rodriguiz, R.M., Huang, X.P., Norval, S., Sassano, M.F., Shin, A.I., Webster, L.A., et al. (2012). Automated design of ligands to polypharmacological profiles. Nature *492*, 215–220.

Bora, A., Avram, S., Ciucanu, I., Raica, M., and Avram, S. (2016). Predictive models for fast and effective profiling of kinase inhibitors. J. Chem. Inf. Model. *56*, 895–905.

Chen, J., Zhang, X., and Fernandez, A. (2007). Molecular basis for specificity in the druggable kinome: sequence-based analysis. Bioinformatics *23*, 563–572.

Christmann-Franck, S., van Westen, G.J., Papadatos, G., Beltran Escudie, F., Roberts, A., Overington, J.P., and Domine, D. (2016). Unprecedently large-scale kinase inhibitor set enabling the accurate prediction of compound-kinase activities: a way toward selective promiscuity by design? J. Chem. Inf. Model. *56*, 1654–1675.

Cichonska, A., Ravikumar, B., Parri, E., Timonen, S., Pahikkala, T., Airola, A., Wennerberg, K., Rousu, J., and Aittokallio, T. (2017). Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. PLoS Comput. Biol. *13*, e1005678.

Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., McCarren, P., Hirschman, J.E., Johnston, S.E., Vrcic, A., Wong, B., Khan, M., et al. (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. Nat. Med. *23*, 405–408.

Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., and Zarrinkar, P.P. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. *29*, 1046–1051.

Drewry, D.H., Wells, C.I., Andrews, D.M., Angell, R., Al-Ali, H., Axtman, A.D., Capuzzi, S.J., Elkins, J.M., Ettmayer, P., Frederiksen, M., et al. (2017). Progress towards a public chemogenomic set for protein kinases and a call for contributions. PLoS One *12*, e0181585.

Elkins, J.M., Fedele, V., Szklarz, M., Abdul Azeez, K.R., Salah, E., Mikolajczyk, J., Romanov, S., Sepetov, N., Huang, X.P., Roth, B.L., et al. (2016). Comprehensive characterization of the published kinase inhibitor set. Nat. Biotechnol. *34*, 95–103.

Fabian, M.A., Biggs, W.H., 3rd, Treiber, D.K., Atteridge, C.E., Azimioara, M.D., Benedetti, M.G., Carter, T.A., Ciceri, P., Edeen, P.T., Floyd, M., et al. (2005). A small molecule-kinase interaction map for clinical kinase inhibitors. Nat. Biotechnol. *23*, 329–336.

Fedorov, O., Muller, S., and Knapp, S. (2010). The (un)targeted cancer kinome. Nat. Chem. Biol. *6*, 166–169.

Fligner, M.A., Verducci, J.S., and Blower, P.E. (2002). A modification of the Jaccard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings. Technometrics *44*, 110–119.

Georgi, V., Schiele, F., Berger, B.T., Steffen, A., Marin Zapata, P.A., Briem, H., Menz, S., Preusse, C., Vasta, J.D., Robers, M.B., et al. (2018). Binding kinetics survey of the drugged kinome. J. Am. Chem. Soc. *140*, 15774–15782.

Geppert, H., Horvath, T., Gartner, T., Wrobel, S., and Bajorath, J. (2008). Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. J. Chem. Inf. Model. *48*, 742–746.

Hanson, S.M., Georghiou, G., Thakur, M.K., Miller, W.T., Rest, J.S., Chodera, J.D., and Seeliger, M.A. (2018). What makes a kinase promiscuous for inhibitors? Cell Chem. Biol. *26*, 1–10.

Harding, S.D., Sharman, J.L., Faccenda, E., Southan, C., Pawson, A.J., Ireland, S., Gray, A.J.G., Bruce, L., Alexander, S.P.H., Anderton, S., et al. (2018). The IUPHAR/BPS Guide to Pharmacology in 2018: updates and expansion to encompass the new guide to Immunopharmacology. Nucleic Acids Res. *46*, D1091–D1106.

Hert, J., Keiser, M.J., Irwin, J.J., Oprea, T.I., and Shoichet, B.K. (2008). Quantifying the relationships among drug classes. J. Chem. Inf. Model. *48*, 755–765.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A., et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc. Natl. Acad. Sci. U S A *107*, 14621–14626.

Jones, L.H., and Bunnage, M.E. (2017). Applications of chemogenomic library screening in drug discovery. Nat. Rev. Drug Discov. *16*, 285–296.

Kang, P.S., and Cho, S.Z. (2006). EUS SVMs: ensemble of under-sampled SVMs for data imbalance problems. Lect. Notes Comput. Sci. *4232*, 837–846.

Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., and Shoichet, B.K. (2007). Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. *25*, 197–206.

Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B., et al. (2009). Predicting new molecular targets for known drugs. Nature *462*, 175–181.

Klaeger, S., Heinzlmeir, S., Wilhelm, M., Polzer, H., Vick, B., Koenig, P.A., Reinecke, M., Ruprecht, B., Petzoldt, S., Meng, C., et al. (2017). The target landscape of clinical kinase drugs. Science *358*, eaan4368.

Kolb, P., Huang, D., Dey, F., and Caflisch, A. (2008). Discovery of kinase inhibitors by high-throughput docking and scoring based on a transferable linear interaction energy model. J. Med. Chem. *51*, 1179–1188.

Koutsoukas, A., Monaghan, K.J., Li, X., and Huan, J. (2017). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. J. Cheminform. *9*, 42.

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. Drug Discov. Today *20*, 318–331.

Lavecchia, A., and Cerchia, C. (2016). In silico methods to address polypharmacology: current status, applications and future perspectives. Drug Discov. Today *21*, 288–298.

Lin, H., Sassano, M.F., Roth, B.L., and Shoichet, B.K. (2013). A pharmacological organization of G protein-coupled receptors. Nat. Methods *10*, 140–146.

Lo, Y.C., Liu, T., Morrissey, K.M., Kakiuchi-Kiyota, S., Johnson, A.R., Broccatelli, F., Zhong, Y., Joshi, A., and Altman, R.B. (2018). Computational analysis of kinase inhibitor selectivity using structural knowledge. Bioinformatics *35*, 235–242.

Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912–1934.

Merget, B., Turk, S., Eid, S., Rippmann, F., and Fulle, S. (2017). Profiling prediction of kinase inhibitors: toward the virtual assay. J. Med. Chem. *60*, 474–485.

Metz, J.T., Johnson, E.F., Soni, N.B., Merta, P.J., Kifle, L., and Hajduk, P.J. (2011). Navigating the kinome. Nat. Chem. Biol. *7*, 200–202.

Miduturu, C.V., Deng, X., Kwiatkowski, N., Yang, W., Brault, L., Filippakopoulos, P., Chung, E., Yang, Q., Schwaller, J., Knapp, S., et al. (2011). High-throughput kinase profiling: a more efficient approach toward the discovery of new kinase inhibitors. Chem. Biol. *18*, 868–879.

Munoz, L. (2017). Non-kinase targets of protein kinase inhibitors. Nat. Rev. Drug Discov. *16*, 424–440.

Orchard, S., Al-Lazikani, B., Bryant, S., Clark, D., Calder, E., Dix, I., Engkvist, O., Forster, M., Gaulton, A., Gilson, M., et al. (2011). Minimum information about a bioactive entity (MIABE). Nat. Rev. Drug Discov. *10*, 661–669.

Ozturk, H., Ozgur, A., and Ozkirimli, E. (2018). DeepDTA: deep drug-target binding affinity prediction. Bioinformatics *34*, i821–i829.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach Learn Res. *12*, 2825–2830.

Pemovska, T., Johnson, E., Kontro, M., Repasky, G.A., Chen, J., Wells, P., Cronin, C.N., McTigue, M., Kallioniemi, O., Porkka, K., et al. (2015). Axitinib

effectively inhibits BCR-ABL1 (T315I) with a distinct binding conformation. Nature *519*, 102–U225.

Ravikumar, B., and Aittokallio, T. (2018). Improving the efficacy-safety balance of polypharmacology in multi-target drug discovery. Expert Opin. Drug Discov. *13*, 179–192.

Rodgers, G., Austin, C., Anderson, J., Pawlyk, A., Colvis, C., Margolis, R., and Baker, J. (2018). Glimmers in illuminating the druggable genome. Nat. Rev. Drug Discov. *17*, 301–302.

Rodriguez-Perez, R., Vogt, M., and Bajorath, J. (2017). Support vector machine classification and regression prioritize different structural features for binary compound activity and potency value prediction. ACS Omega *2*, 6371–6379.

Shannon, H.E., Peters, S.C., and Kingston, A.E. (2005). Anticonvulsant effects of LY456236, a selective mGlu1 receptor antagonist. Neuropharmacology *49* (*Suppl 1*), 188–195.

Sun, J., Wei, Q., Zhou, Y., Wang, J., Liu, Q., and Xu, H. (2017). A systematic analysis of FDA-approved anticancer drugs. BMC Syst. Biol. *11*, 87.

Tang, J., Tanoli, Z.U., Ravikumar, B., Alam, Z., Rebane, A., Vaha-Koskela, M., Peddinti, G., van Adrichem, A.J., Wakkinen, J., Jaiswal, A., et al. (2018). Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. Cell Chem Biol *25*, 224–229.e2.

Tym, J.E., Mitsopoulos, C., Coker, E.A., Razaz, P., Schierz, A.C., Antolin, A.A., and Al-Lazikani, B. (2016). canSAR: an updated cancer research and drug discovery knowledgebase. Nucleic Acids Res. *44*, D938–D943.

UniProt Consortium, T (2018). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *46*, 2699.

Wang, Z., Liang, L., Yin, Z., and Lin, J. (2016a). Improving chemical similarity ensemble approach in target prediction. J. Cheminform. *8*, 20.

Wang, Z., Wu, X., Wang, L., Zhang, J., Liu, J., Song, Z., and Tang, Z. (2016b). Facile and efficient synthesis and biological evaluation of 4-anilinoquinazoline derivatives as EGFR inhibitors. Bioorg. Med. Chem. Lett. *26*, 2589–2593.

Wetzel, S., Klein, K., Renner, S., Rauh, D., Oprea, T.I., Mutzel, P., and Waldmann, H. (2009). Interactive exploration of chemical space with Scaffold Hunter. Nat. Chem. Biol. *5*, 581–583.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082.

Woo, J.H., Shimoni, Y., Yang, W.S., Subramaniam, P., Iyer, A., Nicoletti, P., Rodriguez Martinez, M., Lopez, G., Mattioli, M., Realubit, R., et al. (2015). Elucidating compound mechanism of action by network perturbation analysis. Cell *162*, 441–451.

Zhang, J., Yang, P.L., and Gray, N.S. (2009). Targeting cancer with small molecule kinase inhibitors. Nat. Rev. Cancer *9*, 28–39.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
EGFR	Promega	Cat#V3831; P00533
HCK	Promega	Cat#V1270; P08631
FLT3	Promega	Cat#V4064; P36888
FLT1	Promage	Cat#V3001; P17948
TGFBR2	Promega	Cat#V3931; P37173
RPS6KA5	Promega	Cat#V5092; O75582
CHEK2	Promega	Cat#V4020; O96017
CDK6	Promega	Cat#V4510; Q00534
ABL1	Promega	Cat#V1901; P00519
Critical Commercial Assays		
ADP-Glo Assay kits	Promega	Cat#V6930
Deposited Data		
VirtualKinomeProfiler	This paper	https://virtualkinomeprofiler.fimm.fi/
Software and Algorithms		
GraphPad Prism 7	GraphPad Prism Software, Inc.	https://www.graphpad.com/
Backend development technology: Python 2.7	VirtualKinomeProfiler web-application	https://www.python.org/downloads/release/python-270/
Frontend technology: Jquery 1.11.1, JavaScript	VirtualKinomeProfiler web-application	https://blog.jquery.com/2014/05/01/jquery-1-11-1-and-2-1-1-released/
Virtual Kinome Profiler v1.0 Python v2.7.0 RDKit v2016.09.01	VirtualKinomeProfiler source-code	https://github.com/BalaguruRavikumar/VirtualKinomeProfiler
Figures and Methods: Python, R, D3	This paper	https://www.python.org/downloads/release/python-270/https://www.r-project.org/https://d3js.org/
Kinome Tree	Cell Signaling Technology, Inc.	http://www.kinhub.org/kinmap/

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Tero Aittokallio (tero.aittokallio@helsinki.fi)

METHOD DETAILS

Single-Dose and Dose-Response Assays

Kinase enzyme systems and ADP-Glo Assay kits (Promega) were used to experimentally validate the predictions as previously described with minor modifications (Cichonska et al., 2017). The inhibitors were dissolved into 200nl or 400nl of kinase buffer with 5% DMSO and the kinase was transferred at 400nl or 200nl in the optimization steps or in the inhibitor dose-response screen, respectively. 400nl of substrate/ATP mix was added to attain a kinase reaction volume 1ul. The kinase amounts were optimized on 2-fold dilutions (11 amounts from 40ng to 0.03906ng) at 50 μM ATP concentration with three technical replicates. The ATP concentrations were optimized at 7 concentrations on 2-fold dilutions (from 50 μM to 0.78125 μM), with two technical replicates for each kinase in the presence of positive control inhibitors. The compound dose-response assay was performed at 12 compound concentrations from 12500 nM to 0.0375 nM, with three technical and two biological replicates using the optimized kinase and ATP concentrations. Analysis of the results was performed using GraphPad Prism 7 (GraphPad Software, Inc. California, USA).

$$\%\ Residual\ Activity = \frac{sample - positive\ control}{negative\ control - positive\ control} \times 100 \qquad \text{(Equation 1)}$$

positive control = reaction in absence of kinase
negative control = reaction in absence of compound

Kinome Profiling Dataset: Compilation and Curation

The set of molecular kinases that form the druggable kinome panel in this study was curated from a list of 459 unique kinases manually extracted and compiled from a previous work (Christmann-Franck et al., 2016). The biochemical screening information pertaining to these kinases were retrieved from numerous studies and compiled from existing activity resources, such as ChEMBL, canSAR, and DrugKiNET (http://www.drugkinet.ca/) (Bento et al., 2014; Tym et al., 2016), which includes the interaction estimates (bioactivity values), the compound's structural descriptions (canonical SMILES) and target's sequence information. Since this study focuses on small-molecule kinase inhibitors, the compounds with molecular weight >700 Daltons and with the number of nitrogen or oxygen atoms > 8 were excluded (Keiser et al., 2007; Lin et al., 2013). The resulting compounds' protonation states and valency criteria were satisfied using Open Babel (http://openbabel.org/). To evaluate the performance of the implemented model, compounds that overlap with the previously published kinase chemogenomic set (Drewry et al., 2017; Elkins et al., 2016) were initially held-out during the model development phase. To standardize the response measures, the compound affinity values in the interaction dataset were restricted to dose-response activity end-points (IC_{50}, K_i or K_d). For those compound-kinase interactions that had multiple affinity measures from distinct sources, the activity values were summarized by their geometric mean (Wang et al., 2016a). An activity threshold of 1μM (Christmann-Franck et al., 2016; Lin et al., 2013) was later applied to classify the active and inactive binding classes in the compound-target interaction dataset. To calculate statistically relevant compound-centric estimates, kinase targets with fewer than 10 active ligands and 7 inactive ligands were excluded from the study. The resulting set of 248 kinases (48% of the human kinome) with 74,033 kinase inhibitors and 251,078 compound-kinase interactions form the druggable kinome panel in our study (Figure 1A and Table S1). To the best of our knowledge, the compiled list of annotated KIs, their target interactions, and structural descriptions is the most comprehensive kinome-specific activity resource to date.

Feature Enumeration: Compound Set Similarity and Dissimilarity Estimators

Compound fingerprints (FPs) were enumerated for the pre-determined active and inactive compound sets of each 248 kinase targets using the RDKit chemoinformatics python module (http://www.rdkit.org). FPs were calculated using eight different topological compound representations (Daylight, MACCS, ECFP4/6, Torsion, Atompairs, FCFP4/6) (Figure 1B). Both compound-compound similarity and dissimilarity (1-similarity) estimates were calculated using the Tanimoto coefficient (*Tc*) for MACCS and Daylight FPs and using the Dice index (DI) for the other FPs (Fligner et al., 2002) (Equations 2, 3, 4, and 5). The assumption underlying the implemented compound-centric statistical model is that two kinase targets (Kinases 1 and 2) share a common chemical space if the compounds that actively bind to them are structurally similar and if the active compounds of Kinase 1 are structurally dissimilar to the inactive compound sets of Kinase 2. Therefore, the chemogenomic similarities among the 248 kinase targets were estimated using both their active compound similarity space and their inactive compound dissimilarity space and utilizing all the eight different FP representations (Figure 1B).

Given compounds A and B, let us define;
a = number of bits set in compound A
b = number of bits set in compound B
c = number of bits set common in A and B
S_1 = kinase 1's active or inactive compound set
S_2 = kinase 2's active or inactive compound set
*Tc** = optimal similarity (or dissimilarity) threshold

$$Tc\ (Similarity) = \frac{c}{a+b+c} \quad \text{(Equation 2)}$$

$$Tc\ (Dissimilarity) = 1 - \frac{c}{a+b+c} \quad \text{(Equation 3)}$$

$$DI\ (Similarity) = \frac{c}{\alpha \times a + \beta \times b + c}\ ; \alpha = \beta = 0.5 \quad \text{(Equation 4)}$$

$$DI\ (Dissimilarity) = 1 - \frac{c}{\alpha \times a + \beta \times b + c}\ ; \alpha = \beta = 0.5 \quad \text{(Equation 5)}$$

$$Score_{S1,S2} = \sum_{\substack{x \in S1; y \in S2 \\ Tc(x,y) \geq Tc^*}} Tc(x,y) \quad \text{(and analogous for DI)} \qquad \text{(Equation 6)}$$

Prediction Model: Implementation of an Ensemble SVM Model (eSVM)

The initially held-out kinase chemogenomic screening data (KCGS), consisting both of the published kinase inhibitors screening datasets, PKIS1 (Elkins et al., 2016) and PKIS2 (Drewry et al., 2017) (see Figures S3A and S3B), were used to build and validate the classification model (Figure 1C). The input features were the statistical Z-score estimates obtained from the 16 FPs' similarity and dissimilarity enumerations. Prior to implementing a classification algorithm, the performance of the distinct FP-based statistical features and various baseline models were evaluated in terms of classifying the activity classes in PKIS1 dataset. The evaluation of various baseline models and the implementation of our selected Support Vector Machine (SVM) utilizing the radial basis kernel function (SVM-RBF) were carried out using the scikit-learn machine learning module in python (Pedregosa et al., 2011). The PKIS1 dataset consisting of 54,316 compound-kinase interactions (with 2,821 positives and 51,495 negatives) were split into training and test sets using the standard 80:20 ratio (Figure S3A). The test set consisted of 1,128 interactions (564 positives and 564 negatives), and the training set consisted of 53,188 interactions (2,257 positives and 50,931 negatives). Similar to traditional screening datasets, the ratio of positives to negatives in the PKIS1 dataset was approximately 1:20. To address this class imbalance in the training dataset, we used an ensemble SVM model (eSVM) (Kang and Cho, 2006) as the classification algorithm. To build a generalized SVM model that accounts for the class imbalance and to prevent model overfitting, 23 (50931/2257) ensemble SVMs were generated and subjected to 5-fold cross validations. The various performance metrics, used in evaluating the eSVM model, were employed by averaging the values across all the 23 ensembles (Equations 14, 15, 16, 17, 18, and 19). Using such performance metrics, the model hyperparameters (C and Gamma) were fine-tuned prior to eSVM model implementation. The activity class for a given compound-kinase interaction was scored by normalizing the activity class determined by all the 23 eSVM's, termed as normalized SVM score (Equation 7). The performance of the integrated statistical and eSVM model, in comparison to standard fingerprints (ECFP4, Atompairs and Torsion FPs) in classifying the activity classes of PKIS1 test set was determined by the AUC metric. Both the ECFP4 and Atompairs fingerprints were selected for this comparison as they have been employed in similar classification tasks in the previous studies (Keiser et al., 2007; Lin et al., 2013; Wang et al., 2016a). The proposed integrative analysis framework was further validated using the independent PKIS2 dataset, consisting of 1,057 compound-kinase interactions with 612 positives and 455 negatives (Figure S3B).

Given a compound-kinase interaction i:

$$\text{normalized SVM score}_{(i)} = \frac{PE}{NE} \qquad \text{(Equation 7)}$$

$$\text{prediction score}_{(i)} = \frac{SFP}{NFP} \qquad \text{(Equation 8)}$$

PE = number of ensembles that classify *i* as positive.
NE = total number of ensembles in the SVM model.
SFP = number of FP estimates with significant E-values ($\leq 1 \times 10^{-5}$)
NFP = number of FP estimates in the statistical model

The above-mentioned scoring functions were established to summarize the compound-kinase associations investigated using both the statistical E-value estimates and the SVM models (normalized SVM score and prediction score ranges from 0 to 1). A higher score signifies higher a confidence for the true compound-kinase interaction predictions.

Application of the Model: Compound Library for Virtual Profiling

The integrative computational framework encompassing the developed statistical model and the eSVM classification algorithm was later applied to predict novel insights into compound-kinase interactions (Figures 1A–1D). A compendium of approximately 150,000 compounds from both the repurposing and lead screening libraries were utilized for such classification task (Table S3). These compounds were curated using a protocol similar to that for the kinome dataset preparation, in addition, compounds that are likely to interfere with broad molecular targets were filtered using the PAINS filter (Baell and Holloway, 2010). The resulting compound set was virtually profiled across the 248 kinase targets, where approximately 37 million compound-kinase interactions were classified using the eSVM model. A threshold of $\geq$0.875 for both the normalized SVM and prediction scores were used to select 51 interactions that were later experimentally validated through single-dose and dose-response biochemical assays (Equations 7 and 8). This threshold of $\geq$0.875 was selected based on a single-dose response pilot study, wherein the residual activity of 4 compound-kinase interactions with scores ranging between 0.625 and 1 was initially experimentally

validated (Table S4). Those interactions with normalized SVM score and prediction score $\geq$0.875 were found to be true positives with the % residual activity of <20%, hence the threshold of $\geq$0.875 was used for selecting the 51 predicted interaction for experimental validation.

A Web-Application to Virtually Profile the Druggable Kinome

To promote the wide application of the computational platform and the accompanying data resource, we have implemented the analysis framework, consisting of both the statistical model and eSVM algorithm, as an easy-to-use web-application, termed Virtual Kinome Profiler (https://virtualkinomeprofiler.fimm.fi/). VKP enables end-users to efficiently utilize the platform to systematically classify and prioritize kinome-specific activities of their compounds of interest across a pre-defined set of druggable kinases. The web-application merely requires the structural description (SMILES) of the compound as input, to perform the whole virtual screening process (each compound takes approximately 2.5 h to virtually screen against the selected panel of 248 kinases). In concurrence with high-throughput biochemical screening procedure, the web-application facilitates *in-silico* profiling of compounds across all the kinase targets simultaneously (248 in the present version). However, once more bioactivity data will be collected and standardized for these and other kinases, the kinome coverage and accuracy of the web-application will increase accordingly. In addition to the computational platform, the comprehensive compound sets and their associated kinome-specific activity resource can be downloaded as stand-alone repositories for local use. The output from the web-application includes the statistical estimates from various FPs and the eSVM scores for each of the evaluated kinase-compound interaction, which can be downloaded as a *.csv* or *.pdf* file for further analysis (Figure 1D and Table S7).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical Model: Null Distribution Mean and Standard Deviation

When comparing ligand sets of distinct kinase targets 1 and 2, the statistical significance of the similarity and dissimilarity estimates (Equation 6) was evaluated against a reference null distribution, and the corresponding mean and standard deviations were evaluated at different similarity and dissimilarity thresholds (Tc^* ranging from 0 to 1; step of 0.01). The null distributions were generated for each FP by randomly sampling ligand set sizes from a range of 100 to 1×10^6 product set sizes (e.g. $S_1\times S_2$), observed among the kinase targets in the curated interaction dataset. Such sampling was performed over 10^3 permutations, and the original scores (Equation 6) were calculated for each product set size at the distinct Tanimoto thresholds (99 Tc thresholds). For each FP, the average scores (across 10^3 permutations) for a given Tc captures the relationship of mean and standard deviation with the sampled product set sizes in the curated dataset (see also Figures 2A and 2B). Both the mean (μ) and standard deviation (σ) were non-linearly fit to the sampled product set sizes using the Scipy python module (http://www.scipy.org/) (see Table S2). For a given product set sizes observed among two kinase targets, the expected mean and standard deviations were used to convert the original Scores to Z-scores (Equations 9 and 10).

For each FP at a Tanimoto threshold Tc [i]:

$$Score_{S1,S2,i} = \sum_{\substack{x\in S1; y\in S2 \\ Tc(x,y)\geq Tc[i]}} Tc(x,y) \qquad \text{(Equation 9)}$$

$$Z-Score_{S1,S2,i} = \frac{Score_{S1,S2,i} - \mu(Tc[i])}{\sigma\,(Tc[i])} \qquad \text{(Equation 10)}$$

Statistical Model: Optimal Thresholds and Statistical E-Values

The distribution of Z-scores (Equation 10) estimated for an individual FP obtained at a given Tc threshold corresponds to an Extreme Value Distribution (EVD) (Figure 1B). This is similar to the EVD observed in the standard BLAST algorithm (Altschul et al., 1990) implemented for sequence similarity searches. The estimated EVD gives the likelihood of observing a similar or better score at random. Hence, each of the 99 *Z* score distributions (for each FP) was fitted to an empirical generalized extreme value Gumbel distribution (where right tail corresponds to active similarities; and left tail to inactive dissimilarities), thereby providing the probability of obtaining the observed Z-scores at random (E-values). An E-value statistic gives the balanced likelihood of observing the score by a random chance by accounting for the disproportionate compound set sizes related to each kinase targets. Each *Z* score (z) distribution across the 99 Tc thresholds was considered as an independent statistical score and the optimal Tc threshold (Tc^*) (Equation 6), the threshold at which a *Z* score distribution best fits an empirical EVD, was defined using the Akaike Information Criterion (AIC) (see Figure 2C). In other words, AIC was used as a goodness-of-fit estimator as the number of parameters (location and scale) was constant (k = 2) across the different statistical models. Using the optimal Tanimoto thresholds (Tc^*) (Table S2; Figure 2D) for similarity and dissimilarity estimates across each fingerprint, the expected mean, standard deviation, and the corresponding Z-scores were calculated (Equations 9 and 10). The E-value for an observed *Z* score of z is calculated based on a cumulative Gumbel distribution, with zero mean and unit standard deviation:

$$P(Z>z) = 1 - e^{-e^{-\frac{\pi}{\sqrt{6}}z - \Gamma'(1)}}, \quad \text{(Equation 11)}$$

where π is *pi* ($\approx$3.14159), and $\Gamma'(1)$ is the Euler-Mascheroni constant ($\approx$0.57721).

For $z > 28$, the computation exceeds the numerical precision of the programming languages, hence a Taylor series approximation was used:

$$P(Z>z) = -\left(x + \frac{x^2}{2} + \frac{x^3}{6}\right), \quad \text{(Equation 12)}$$

where $x = -e^{-\frac{\pi}{\sqrt{6}}z - \Gamma'(1)}$.The corresponding E-value is calculated as:

$$E(z) = P(z)N, \quad \text{(Equation 13)}$$

where N is the number of dataset searches.

The steps detailed in the above statistical module of the analysis framework implement significant improvements over the similarity ensemble approach (Keiser et al., 2007; Lin et al., 2013; Wang et al., 2016a). These include, for instance, adopting a kinase-specific model rather than an all-inclusive framework, evaluation of both the active and inactive ligand set similarity and dissimilarity estimates from eight distinct molecular FPs, making use of data-specific optimal FP thresholds, and finally the implementation of an efficient eSVM classification algorithm to consolidate the non-redundant information from the 16 features.

Evaluation: Performance Measures for Classification Model Accuracy

The binary classification performance of the various machine learning classifiers were evaluated using statistical performance measures calculated based on a confusion matrix. In case of the ensemble SVM model, the results of the performance measures were average across all the 23 ensembles.

Confusion matrix:

Observed/Predicted	0	1
0	TN	FP
1	FN	TP

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{(Equation 14)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{(Equation 15)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{(Equation 16)}$$

$$False\ Discovery\ Rate = \frac{FP}{TP+FP} \quad \text{(Equation 17)}$$

$$F-measure = 2 \times \frac{Precision \times Recal}{Precision + Recall} \quad \text{(Equation 18)}$$

$$F\beta = (1+\beta)^2 \times \frac{Precision \times Recal}{\beta^2 \times Precision + Recall}; \beta = 0.5 \quad \text{(Equation 19)}$$

Visualization: Chemogenomic Clustering and Differential Heatmaps

The statistical E-value estimates portray the strength of chemogenomic similarities or dissimilarities observed among the 248 kinases in our druggable kinome panel. Each of the target-target similarity/dissimilarity matrices (a total of 16 matrices: 8 FPs used in active ligand set similarity estimates and 8 FPs used in inactive ligand set dissimilarity estimates) were converted to a binary matrix using an

E-value threshold of 1×10^{-10}. Each such binary matrix was converted to a distance matrix by calculating the cosine distance among the 248-binary target vectors. The chemogenomic clustering of these 16 distinct target-target distance matrices was carried out using an unsupervised agglomerative hierarchical clustering algorithm, namely Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Lin et al., 2013). The correlation among these clustering solutions and their subcluster diversity were enumerated using cophenetic correlation and adjusted Rand index. A similar clustering protocol was also implemented to visualize the target-target sequence similarity E-values obtained from the BLASTp search algorithm. The E-value matrices obtained from both the ECFP4 fingerprint and the sequence similarity algorithm were log-transformed and their differences, highlighting the target-target similarities from these orthogonal approaches, were depicted as a differential heatmap.

DATA AND CODE AVAILABILITY

The VirtualKinomeProfiler web-application and the accompanying data supporting the current study are hosted on the FIMM server and are freely available through the VKP website: https://virtualkinomeprofiler.fimm.fi/. The codes implementing the chemogenomic analysis and prediction framework are available under the Mozilla Public License 2.0 (https://github.com/BalaguruRavikumar/VirtualKinomeProfiler).